INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions

Hao Zhang[1], *Ke Tan*[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

{zhang.6720, tan.650, wang.77}@osu.edu

Abstract

We formulate acoustic echo and noise cancellation jointly as deep learning based speech separation, where near-end speech is separated from a single microphone recording and sent to the far end. We propose a causal system to address this problem, which incorporates a convolutional recurrent network (CRN) and a recurrent network with long short-term memory (LSTM). The system is trained to estimate the real and imaginary spectrograms of near-end speech and detect the activity of near-end speech from the microphone signal and far-end signal. Subsequently, the estimated real and imaginary spectrograms are used to separate the near-end signal, hence removing echo and noise. The trained near-end speech detector is employed to further suppress residual echo and noise. Evaluation results show that the proposed method effectively removes acoustic echo and background noise in the presence of nonlinear distortions for both simulated and measured room impulse responses (RIRs). Additionally, the proposed method generalizes well to untrained noises, RIRs and speakers.

Index Terms: Acoustic echo cancellation, supervised speech separation, deep learning, complex spectral mapping, nonlinear distortion

1. Introduction

Acoustic echo arises when a loudspeaker and a microphone are coupled in a communication system such that the microphone picks up the loudspeaker signal plus its reverberation. If not properly handled, a user at the far end of the system hears his or her own voice delayed by the round trip time of the system (i.e. an echo), mixed with the target speech signal from the near end. Traditional acoustic echo cancellation (AEC) works by identifying a room impulse response using adaptive algorithms [1]. Many algorithms have been proposed in the literature [1–4]. However, the performance of these algorithms is limited in the presence of double-talk (both near-end and far-end speakers are talking), background noise (especially non-stationary noises), and nonlinear distortions.

Typical approaches to the double-talk problem are to use double-talk detectors [5] or double-talk-robust AEC algorithms [6]. In a noisy environment, post-filtering [7, 8], Kalman filtering [9], and spectral modification based acoustic echo suppression (AES) algorithms [10–12] are usually used. Nonlinear distortions are introduced mainly due to the poor quality of electronic devices such as amplifiers and loudspeakers. Traditional AEC algorithms are essentially linear systems, which suffer from nonlinear distortions. In order to address the nonlinear distortion problem, algorithms such as adaptive Volterra filters [13] and functional link adaptive filters [14] have been recently investigated to model the nonlinearity of AEC system.

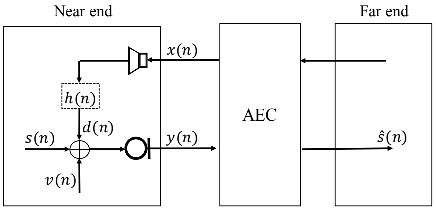

Figure 1: *Diagram of an acoustic echo scenario.*

Deep neural networks (DNNs) are fundamentally capable of modeling complicated nonlinear relationships, and they are expected to provide a compelling alternative to linear AEC algorithms. Early work [15, 16] used a cascaded time-delay feedforward neural network (TDNN) to model the nonlinearity of the acoustic channel. In a recent study [17], a DNN was used as a residual echo suppression to suppress the nonlinear components of the echo. More recently we formulated AEC as a supervised speech separation problem and proposed a deep learning based AEC method [18]. Compared with traditional AEC algorithms, deep learning based methods avoid the need to perform double-talk detection or post filtering. However, previous deep learning based methods are trained in a noise- and RIR-dependent way with limited robustness.

The ultimate goal of AEC in a noisy environment is to completely cancel echo and background noise and transmit only near-end speech to the far end [18, 19]. From the speech separation point of view, we address this problem as a supervised speech separation problem [20] where the near-end speech signal is the target source to be separated from the microphone recording. Deep learning has yielded great advances in speech separation [20–22], and will likely play an important role in addressing AEC challenges.

A recent study [23] shows that accurate phase estimation can lead to considerable improvements in speech quality. In this paper, a CRN is trained for complex spectral mapping [24], which estimates the real and imaginary spectrograms of near-end speech. Hence, it is capable of enhancing both magnitude and phase responses simultaneously. Motivated by the potential of residual echo suppression in removing the residual echo at the output of AEC, a near-end speech detector (NSD) is estimated with an LSTM network to further suppress residual echo and noise. The proposed system is trained in a noise- and RIR-independent way, and can generalize to untrained noises and RIRs.

The remainder of this paper is organized as follows. Section 2 presents the proposed method. Evaluation metrics and experimental results are shown in Section 3. Section 4 concludes the paper.

http://dx.doi.org/10.21437/Interspeech.2019-2651

2. Proposed method

The acoustic signal model is shown in Fig. 1. The microphone signal $y(n)$ is a mixture of echo $d(n)$, near-end speech $s(n)$, and background noise $v(n)$:

$$y(n) = d(n) + s(n) + v(n) \tag{1}$$

where n indexes a time sample, and echo is generated by convolving a loudspeaker signal with an RIR. The echo $d(n)$ is typically a linear or nonlinear transform of the far-end signal $x(n)$, as illustrated in Fig. 1. We formulate AEC as a supervised speech separation problem. As shown in Fig. 2, the overall approach is to estimate the real and imaginary spectrograms of near-end speech as well as the NSD from $y(n)$ and $x(n)$ to suppress the acoustic echo and background noise, and isolate the embedded near-end speech.

2.1. Feature extraction

The CRN takes the real and imaginary spectrograms of input signals ($y(n)$ and $x(n)$), while LSTM$_2$ takes the magnitude spectrograms of them as input features. The input signals, sampled at 16 kHz, are divided into 20-ms frames with a 10-ms overlap between consecutive frames. Then a 320-point short time Fourier transform (STFT) is applied to each time frame to produce the real, imaginary and magnitude spectra ($*_\mathrm{r}$, $*_\mathrm{i}$ and $*_\mathrm{m}$) of input signals.

2.2. Training targets

We explore two training targets in this study:

- **Complex spectrum of near-end speech** [20]: The real and imaginary spectrograms of near-end speech are used as the training targets of the CRN. Let $S_\mathrm{r}(m,c)$ and $S_\mathrm{i}(m,c)$ denote the targets within a T-F unit at time m and frequency c, respectively. Different from magnitude spectral mapping/masking based methods that use noisy phase for waveform resynthesis, complex spectral mapping can enhance both magnitude and phase responses through supervised learning and thus further improve speech quality.
- **Near-end speech detector**: An NSD can be regarded as a frame-level binary mask that detects the activity of near-end speech. If no near-end speech is present at frame m, $NSD(m) = 0$; otherwise, $NSD(m) = 1$:

$$NSD(m) = \begin{cases} 1, & \text{if } \max_c |S(m,c)| > 0 \\ 0, & \text{else} \end{cases} \tag{2}$$

The NSD estimated by LSTM$_2$ is applied to the estimated complex spectrogram to suppress residual echo and noise at the frames without the presence of near-end speech while maintaining near-end speech estimated by the CRN.

2.3. Learning machines

The proposed system consists of two components. First, a CRN is employed to predict the complex spectrum of near-end speech [24]. It is an encoder-decoder architecture as depicted in Fig. 2. Specifically, the encoder and decoder comprise five convolutional layers and five deconvolutional layers, respectively. Between them is a two-layer LSTM with a group strategy [25], where the group number is set to 2. A detailed description of the CRN architecture is provided in [24] except that our CRN has four input channels, which correspond to the real and imaginary spectra of the microphone signal (Y_r, Y_i) and the far-end signal (X_r, X_i), respectively. An LSTM, LSTM$_2$, is used to predict the NSD from the magnitude spectrograms of input signals (Y_m, X_m). LSTM$_2$ has four hidden layers with 300 units in each layer. The output layer is a fully-connected layer. Sigmoid function is used as the activation function in the output layer. The AMSGrad optimizer [26] and the mean squared error (MSE) cost function are used to train both networks. The networks are trained for 30 epochs with a learning rate of 0.001. The minibatch size is set to 16 at the utterance level.

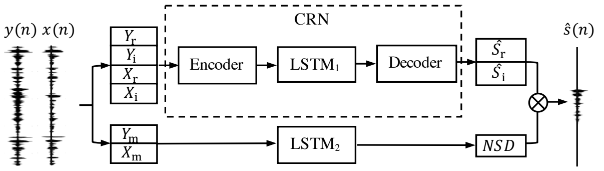

Figure 2: *Diagram of the proposed system.*

2.4. Signal resynthesis

The output of the CRN is an estimate of the complex spectrogram of near-end speech:

$$\hat{S}(m,c) = \hat{S}_\mathrm{r}(m,c) + i\hat{S}_\mathrm{i}(m,c) \tag{3}$$

where i denotes the imaginary unit. When applying the estimated NSD, the estimate $\hat{S}(m,c)$ can be modified by element-wise multiplications:

$$\hat{S}(m,c)_\mathrm{NSD} = NSD(m) \cdot \hat{S}(m,c) \tag{4}$$

The estimated complex spectrogram and the modified complex spectrogram of near-end speech are fed into the inverse short time Fourier transform based resynthesizer to derive the time-domain estimated near-end speech signals.

Note that if the NSD is estimated accurately, $\hat{S}(m,c)_\mathrm{NSD}$ in the the single-talk period (with the far-end signal only and no near-end speech) should be all zeros. In other words, residual echo and noise in the single-talk period of $\hat{S}(m,c)$ is completely removed. Thus, the echo return loss enhancement (ERLE) in this period can be improved to infinity.

3. Experimental results

3.1. Performance metrics

The performance of the proposed method is evaluated in terms of ERLE [2] for single-talk periods and perceptual evaluation of speech quality (PESQ) [27] for double-talk periods. In this study, ERLE is defined as

$$\mathrm{ERLE} = 10\log_{10}\left[\sum_n y^2(n) / \sum_n \hat{s}^2(n)\right] \tag{5}$$

This variant of ERLE is widely used in the literature [9–12] for assessing masks related AEC systems in the presence of background noise. It reflects the integrated echo and noise attenuation achieved by systems.

3.2. Experiment setting

The TIMIT dataset [28] is used in the situations with double-talk, background noise, and nonlinear distortions. To investigate speaker generalization, we randomly choose 100 pairs of speakers (40 pairs of male-female, 30 pairs of male-male, and 30 pairs of female-female) from the 630 speakers in the TIMIT dataset as the near-end and far-end speakers, respectively. Out of the ten utterances of each speaker, seven utterances are randomly chosen to create training mixtures, and the three remaining utterances are used to create test mixtures. Three randomly chosen utterances from a far-end speaker are concatenated to generate a far-end signal. A randomly chosen utterance from a near-end speaker is extended to the same length as that of the far-end signal by zerso-padding both in the beginning and in the end, where the number of leading zeros is random.

To achieve a noise-independent model, we use 10000 noises from a sound effect library (http://www.sound-ideas.com) for the training mixtures. An oproom (operational room) noise from NOISEX-92 dataset [29], a babble noise from the Auditec CD (http://www.auditec.com), and a white noise are used for the test mixtures. The RIRs are generated using the image method [30]. To investigate RIRs generalization, we simulate 20 different rooms of size $a \times b \times c$ m for training mixtures, where $a = [4, 6, 8, 10], b = [5, 7, 9, 11, 13], c = 3$. We randomly choose ten positions in each room with fixed microphone-loudspeaker (M-L) distance (1 m) to generate the RIRs. The length of the RIRs is set to 512, the reverberation time (T_{60}) is randomly selected from $\{0.2, 0.3, 0.4\}$ s. Therefore, in total 200 RIRs are created for training mixtures. For test mixtures, we use both simulated and real measured RIRs. Two of them (RIR$_1$ and RIR$_2$) are generated by the image method with M-L distance of 1 m and T_{60} of 0.2 s. The simulation room sizes, $3 \times 4 \times 3$ m and $11 \times 14 \times 3$ m, are different from the 20 rooms used for training mixtures. The other two RIRs (RIR$_3$ and RIR$_4$) are selected from the Aachen impulse response database [31]. They are measured in a meeting room of size $8 \times 5 \times 3.1$ m. The T_{60} is 0.23 s, and M-L distances of them are 1.45 m and 2.25 m, respectively. Note that RIR$_3$ and RIR$_4$ are measured from far-end signals and microphone signals. That is to say, they are correlated with the transfer functions between far-end signals and loudspeaker signals.

We create 20000 training mixtures and 300 test mixtures. Each training mixture is created by first convolving a randomly chosen loudspeaker signal (or far-end signal for conditions without nonlinear distortions) with a randomly chosen RIR from the 200 training RIRs to generate an echo. Then a randomly chosen near-end speech is mixed with the echo at a signal-to-echo ratio (SER) randomly chosen from $\{-6, -3, 0, 3, 6\}$ dB. Finally, a random cut from the 10000 noises is added to the mixture at a signal-to-noise ratio (SNR) randomly chosen from $\{8, 10, 12, 14\}$ dB. The SER and SNR, which are evaluated during double-talk periods, are defined as:

$$\mathrm{SER} = 10 \log_{10} \left[\textstyle\sum_n s^2(n) / \sum_n d^2(n) \right] \tag{6}$$

$$\mathrm{SNR} = 10 \log_{10} \left[\textstyle\sum_n s^2(n) / \sum_n v^2(n) \right] \tag{7}$$

Test mixtures are created similarly but using different utterances, noises, RIRs, SERs and SNRs.

3.3. Performance in double-talk and background noise situations

We first compare the proposed method with some traditional methods in the scenarios with double-talk and background noise. The joint-optimized normalized least mean square (JO-NLMS) algorithm is a recently proposed double-talk-robust algorithm that is developed in the context of a state-variable model and tries to minimize the system misalignment [6]. The parameters for JO-NLMS and AES are set to the values given in [6] and [10] respectively. Since JO-NLMS alone is not capable of handling background noise, a post-filter (PF) [8] is employed to suppress noises at the output of it. The two forgetting factors of the post-filter are set to 0.99.

Table 1: *Performance in the presence of double-talk and babble noise with 3.5 dB SER, 10 dB SNR.*

	ERLE		PESQ	
	RIR$_1$	RIR$_2$	RIR$_1$	RIR$_2$
Unprocessed	-	-	2.00	1.99
JO-NLMS	7.22	7.14	2.37	2.37
JO-NLMS-PF	15.94	15.20	2.47	2.46
AES	7.69	7.65	2.32	2.32
CRN	33.78	37.72	**2.62**	**2.86**
CRN-NSD	**Inf (76.33%)** **43.27 (23.67%)**	**Inf (99%)** **58.73 (1%)**	2.54	2.77

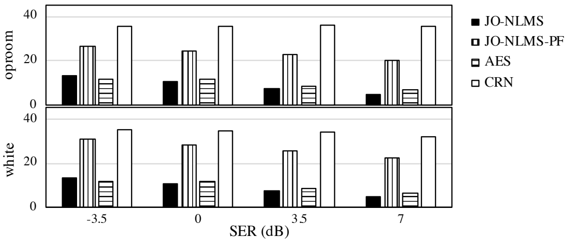

Figure 3: *ERLE values at different SER levels with RIR$_1$.*

Table 1 shows the average ERLE and PESQ values of 300 test mixtures in the presence of double-talk and babble noise with different RIRs. ERLE for JO-NLMS based methods are the steady-state results. In general, the proposed CRN method outperforms conventional methods, especially in terms of ERLE. Furthermore, when combined with NSD (CRN-NSD), ERLE of most test mixtures during single-talk periods can be improved to infinity. As it was mentioned previously, we used an utterance-level ERLE that is defined as the ratio between the sum of microphone signal energy and that of the output signal during the whole single-talk periods. The infinity here means that residual echo and noise for all single-talk time frames are completely removed. Note that the estimation of NSDs for some test mixtures may not be accurate enough. Hence, ERLE of some test mixtures are not improved to infinity. The numbers in the parentheses after "Inf" show the percentage of test mixtures that obtained infinity ERLE. The other two values show the ERLE and percentage of test utterances that are not improved to infinity. Take the results of CRN-NSD with RIR$_1$ as an example. Of the 300 test samples, ERLE of 229 samples are improved to infinity, and the average ERLE of the 71 remaining samples is improved to 43.27 dB. We also observe that the improvement of CRN-NSD in terms of ERLE is at the cost of an acceptable reduction in PESQ. Besides, the proposed method can be generalized to untrained RIRs (RIR$_1$, RIR$_2$). The comparison results in different background noises and SERs are given in Fig. 3. The proposed method consistently outperforms the conventional methods, and the performance generalizes well to untrained noises and SERs.

3.4. Performance in double-talk, background noise and nonlinear distortions situations

The nonlinear distortions introduced by a power amplifier and a loudspeaker are simulated by following steps [17]. First, a hard

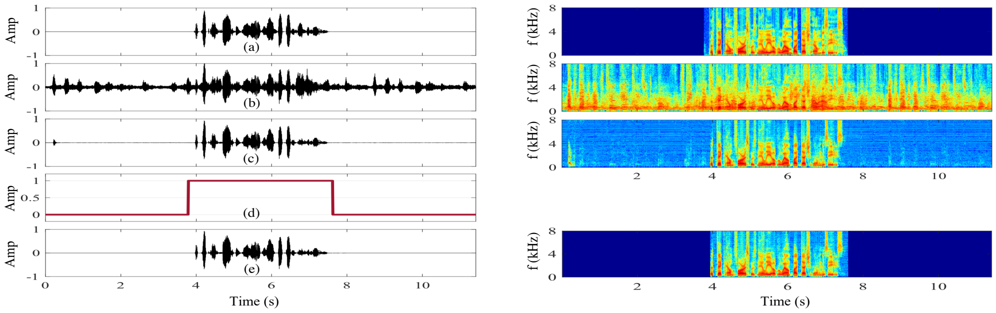

Figure 4: *Waveforms and spectrograms with 3.5 dB SER, 10 dB SNR (babble noise) and nonlinear distortions (RIR$_1$): (a) near-end speech , (b) microphone signal, (c) CRN estimated near-end speech, (e) CRN-NSD estimated near-end speech, and (d) estimated NSD.*

Table 2: *Performance in the double-talk, background noise and nonlinear distortions situations with 3.5 dB SER, 10 dB SNR.*

		RIR$_1$		
		Babble	Oproom	White
ERLE	AES-DNN	19.28	24.16	22.79
	LSTM	43.02	38.77	31.78
	CRN	35.90	32.89	29.46
	CRN-NSD	**Inf (80.00%)** **46.14 (20.00%)**	**Inf (70.67%)** **42.35 (29.33%)**	**Inf (64.67%)** **34.62 (34.33%)**
PESQ	None	1.96	1.97	1.90
	AES-DNN	2.33	2.38	2.22
	LSTM	2.26	2.28	2.24
	CRN	**2.53**	**2.56**	**2.58**
	CRN-NSD	2.45	2.49	2.51

		RIR$_3$		
		Babble	Oproom	White
ERLE	AES-DNN	18.91	24.31	22.69
	LSTM	50.01	47.47	48.42
	CRN	37.15	35.41	35.84
	CRN-NSD	**Inf (95.00%)** **52.78 (5.00%)**	**Inf (84.00%)** **54.68 (16.00%)**	**Inf (96.33%)** **53.01 (3.67%)**
PESQ	None	1.97	1.99	1.90
	AES-DNN	2.34	2.39	2.24
	LSTM	2.32	2.40	2.36
	CRN	**2.48**	**2.56**	**2.50**
	CRN-NSD	2.40	2.50	2.51

clipping [32] is applied to each far-end signal to simulate the characteristic of a power amplifier:

$$x_{\mathtt{hard}}(n) = \begin{cases} -x_{\mathtt{max}} & x(n) < -x_{\mathtt{max}} \\ x(n) & |x(n)| \leq x_{\mathtt{max}} \\ x_{\mathtt{max}} & x(n) > x_{\mathtt{max}} \end{cases} \tag{8}$$

where $x_{\mathtt{max}}$ is set to 0.8 as the maximum amplitude of $|x(n)|$. Then a memoryless sigmoidal nonlinearity [14] is applied to the clipped signal to simulate an asymmetric loudspeaker distortion:

$$x_{\mathtt{NL}}(n) = \gamma\left(\frac{2}{1+\exp(-a \cdot b(n))} - 1\right) \tag{9}$$

where $b(n) = 1.5 \times x_{\mathtt{hard}}(n) - 0.3 \times x^2_{\mathtt{hard}}(n)$. The sigmoid gain γ is set to 4. The sigmoid slope a is set to 4 if $b(n) > 0$ and 0.5 otherwise. Finally, a loudspeaker signal, $x_{\mathtt{NL}}$, is convolved with an RIR to generate an echo with nonlinear distortions.

Waveforms and spectrograms in Fig. 4 illustrate an echo cancellation example of the proposed method, where 'Amp' stands for amplitude. The CRN based method can remove most of the echo and noise in the microphone signal. However, it is obvious that there still exists some amount of residual echo and noise, which can be completely removed by using NSD based residual echo suppression.

We compared the proposed method with a DNN based residual echo suppression method [17] and an LSTM based method [18] (we replace the BLSTM by a unidirectional LSTM). In [17], AES [10] was used for preprocessing and a DNN was employed to remove the residual echo. The parameters for AES-DNN are set to the values given in [17]. The comparison results are given in Table 2. It is evident that all of these deep learning based methods are capable of suppressing echoes in the presence of nonlinear distortions. The proposed CRN-NSD method outperforms the other two methods in most of the cases. Note that the LSTM outperforms the CRN in terms of ERLE. An explanation for this is that the target of LSTM based method is a ratio mask, of which the value range is $[0, 1]$. Generally, the estimation of a ratio mask can be easier and more accurate than directly estimating complex spectrogram. The CRN employs complex spectral mapping which enhances the magnitude and phase responses simultaneously, and hence it yields significantly higher PESQ values than the LSTM.

Table 3: *Performance under echo path change and untrained speakers conditions with 3.5 dB SER,10 dB SNR (babble noise).*

	RIR		None	CRN	CRN-NSD
Echo path change	RIR$_3$, RIR$_4$	ERLE	-	35.43	Inf (85.67%) | 55.86 (14.33%)
		PESQ	1.99	2.43	2.35
Untrained speakers	RIR$_1$	ERLE	-	36.51	Inf (79%) | 47.37 (21%)
		PESQ	1.95	2.53	2.46

Table 3 shows the behavior of the proposed method when the echo path is changed and the test speakers are untrained. The echo path change is simulated by toggling between RIR$_3$ and RIR$_4$ every 1.5 seconds for each test mixture. To create the untrained test mixtures, we randomly select 10 pairs of untrained speakers from the 430 remaining speakers of TIMIT dataset and create 100 test mixtures. The results in this table indicate high robustness of the proposed method.

4. Conclusion

In this paper, we propose a complex spectral mapping based system to address the integrated echo and noise cancellation problem with nonlinear distortions. The performance of the proposed method is further improved by estimating a near-end speech detector. Evaluations show that the proposed system is effective for removing echo and noise for untrained noises as well as untrained simulated and real measured RIRs, and it substantially outperforms previous techniques.

5. Acknowledgements

This research started while the first author was interning with Elevoc Technology, and it was supported in part by two NIDCD grants (R01 DC012048 and R01 DC015521) and the Ohio Supercomputer Center.

6. References

[1] J. Benesty, T. Gänsler, D. Morgan, M. Sondhi, S. Gay *et al.*, *Advances in network and acoustic echo cancellation*. Springer, 2001.

[2] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic press library in signal processing: image, video processing and analysis, hardware, audio, acoustic and speech Processing*. Academic Press, 2014.

[3] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*. John Wiley & Sons, 2005, vol. 40.

[4] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, *A perspective on stereophonic acoustic echo cancellation*. Springer Science & Business Media, 2011, vol. 4.

[5] D. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Transactions on Communications*, vol. 26, no. 5, pp. 647–653, 1978.

[6] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant, "An overview on optimized nlms algorithms for acoustic echo cancellation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 97, 2015.

[7] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1997, pp. 307–310.

[8] F. Ykhlef and H. Ykhlef, "A post-filter for acoustic echo cancellation in frequency domain," in *Second World Conference on Complex Systems*. IEEE, 2014, pp. 446–450.

[9] K. Nathwani, "Joint acoustic echo and noise cancellation using spectral domain kalman filtering in double-talk scenario," in *International Workshop on Acoustic Signal Enhancement*. IEEE, 2018, pp. 1–330.

[10] F. Yang, M. Wu, and J. Yang, "Stereophonic acoustic echo suppression based on wiener filter in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 227–230, 2012.

[11] Y. S. Park and J. H. Chang, "Frequency domain acoustic echo suppression based on soft decision," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 53–56, 2009.

[12] Y. Tong and Y. Gu, "Acoustic echo suppression based on speech presence probability," in *IEEE International Conference on Digital Signal Processing*. IEEE, 2016, pp. 35–38.

[13] A. Stenger, L. Trautmann, and R. Rabenstein, "Nonlinear acoustic echo cancellation with 2nd order adaptive volterra filters," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1999, pp. 877–880.

[14] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-Garcia, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, 2013.

[15] A. N. Birkett and R. A. Goubran, "Acoustic echo cancellation using nlms-neural network structures," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 5. IEEE, 1995, pp. 3035–3038.

[16] ——, "Nonlinear loudspeaker compensation for hands free acoustic echo cancellation," *Electronics Letters*, vol. 32, no. 12, pp. 1063–1064, 1996.

[17] C. M. Lee, J. W. Shin, and N. S. Kim, "Dnn-based residual echo suppression," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] H. Zhang and D. L. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," *Proceedings of INTERSPEECH*, pp. 3239–3243, 2018.

[19] J. M. Portillo, "Deep Learning applied to Acoustic Echo Cancellation," Master's thesis, Aalborg University, 2017.

[20] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[21] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[22] M. Delfarah and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.

[23] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[24] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[25] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T. Liu, "Efficient sequence learning with group recurrent networks," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2018, pp. 799–808.

[26] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2001, pp. 749–752.

[28] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.

[29] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[30] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[31] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *International Conference on Digital Signal Processing*. IEEE, 2009, pp. 1–5.

[32] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 2065–2079, 2012.